Multimedia Systems (2006)
DOI 10.1007/s00530-006-0024-2

REGULAR PAPER

Faisal I. Bashir · Ashfaq A. Khokhar · Dan Schonfeld

View-invariant motion trajectory-based activity classification and recognition

© Springer-Verlag 2006

Abstract Motion trajectories provide rich spatio-temporal information about an object's activity. The trajectory information can be obtained using a tracking algorithm on data streams available from a range of devices including motion sensors, video cameras, haptic devices, etc. Developing view-invariant activity recognition algorithms based on this high dimensional cue is an extremely challenging task. This paper presents efficient activity recognition algorithms using novel view-invariant representation of trajectories. Towards this end, we derive two Affine-invariant representations for motion trajectories based on *curvature scale space* (CSS) and *centroid distance function* (CDF). The properties of these schemes facilitate the design of efficient recognition algorithms based on hidden Markov models (HMMs). In the CSS-based representation, maxima of curvature zero crossings at increasing levels of smoothness are extracted to mark the location and extent of concavities in the curvature. The sequences of these CSS maxima are then modeled by continuous density (HMMs). For the case of CDF, we first segment the trajectory into *subtrajectories* using CDF-based representation. These subtrajectories are then represented by their Principal Component Analysis (PCA) coefficients. The sequences of these PCA coefficients from subtrajectories are then modeled by continuous density hidden Markov models (HMMs). Different classes of object motions are modeled by one Continuous HMM per class where state PDFs are represented by GMMs. Experiments using a database of around 1750 complex trajectories (obtained from UCI-KDD data archives) subdivided into five different classes are reported.

Keywords Affine-invariant trajectory descriptors · Trajectory modeling · Activity recognition · Hidden Markov models · Centroid distance function · Curvature scale space

F. I. Bashir (✉) · A. A. Khokhar · D. Schonfeld
Department of Electrical and Computer Engineering,
University of Illinois at Chicago, 851 S. Morgan Street,
Chicago, IL 60607
E-mail: {fbashir, ashfaq, ds}@ece.uic.edu

1 Introduction

Object motion trajectory-based analysis and recognition has gained significant interest in scientific circles lately. Examples of the motion trajectory include tracking results from video trackers, sign language data measurements gathered from wired glove interfaces fitted with sensors, Global Positioning System (GPS) coordinates of satellite phones, cars using Car Navigation Systems (CNS), animal mobility experiments, etc. This spatio-temporal data embodies semantically rich information about the behavior of the object of interest, the action performed and the interaction among groups of objects [10]. Novel applications employing analysis of motion trajectory are emerging due to enhanced interest in homeland security as well as due to prevalence of multimedia gadgets in commercial and scientific endeavors. For example, in sign and gesture recognition, the signer moves his hands in specific pattern for a particular word. In sports video trajectory analysis and understanding can assist the players, coaches and sports analysts with strategies used on the field based on the motion patterns of players and their mutual interaction. Another important area is automatic video surveillance which is used, for example, in real-time observation of people and vehicles, in a busy environment, leading to a description of actions and mutual interactions. Psychological studies have shown that human beings can routinely discriminate and recognize this kind of object motion using motion pattern, even in large viewing distances or poor visibility conditions; whereas, other cues such as clothes, appearance, or hair style tend to vanish at large distances or poor visibility conditions [23].

We observe that developing high-accuracy activity classification and recognition algorithms in multiple view situations is still an extremely challenging task. The object trajectory is typically modeled as a sequence of consecutive locations of the object on a coordinate system resulting in a vector in 2-D or 3-D Euclidean space. An object trajectory captured from different view-points leads to entirely different representations. Multiple representations of an object motion captured from different view points are related by perspective transformation exactly, and by generic affine transformation approximately. The set of affine transformations includes scaling, rotation, translation and shear. To satisfy the view independent requirement, the trajectory data has to be represented in an affine invariant feature space.

This paper presents efficient object activity recognition algorithms based on novel view-invariant representation of trajectories. Since we are addressing the problem of classification of activities based on trajectory data only, we cannot ensure the semantic interpretation of all actions resulting from those trajectories. This problem can be addressed by the addition of other cues along with motion trajectories in the design of activity recognition systems. We derive two Affine-invariant representations for motion trajectories based on *curvature scale space* (CSS) and *centroid distance function* (CDF). These feature spaces allow affine invariant representation of trajectories as temporal sequences of subtrajectories analogous to the characterization of words as a sequence of phonemes. We subsequently capture the interactions among subtrajectories using ergodic Hidden Markov Models (HMM). All HMM parameters including the topology are learnt from training datasets. The performance of classification of motion trajectories using CSS- and CDF-based view-invariant representations is analyzed and computer simulation experiments are reported. The comparisons are performed in terms of *receiver operating characteristics* (ROC) as well as accuracy values. Experiments are conducted on the Australian Sign Language (ASL) data set obtained from University of California at Irvine's Knowledge Discovery in Databases archive [18]. The ASL dataset used in our experiments has trajectories for five word classes as signed by five signers of varying skill levels. Each word class has 69 trajectories for a total of 345 original samples. The training is performed on half of these samples, while for testing an affined database containing rotated version of these trajectories at five different angles is generated. Thus from the original set of 345 trajectories, our training set contains 170 trajectories while the test set contains 1725 trajectories. Our results show that overall, CDF-based representations perform better in terms of ROC and accuracy values.

The remaining sections of this paper are organized as follows: Section 2 surveys related work on trajectory modeling and view-invariant feature spaces. Section 3 briefly describes the two view-invariant feature spaces used in this presentation. Section 4 presents the HMM-based view-invariant action recognition using CSS and CDF-based representations. Section 5 provides a comparison of the above methods in terms of their retrieval performance. The details of the data sets used and the experiments conducted are also provided in Sect. 5. Section 6 provides a discussion of the results of the computer experiments. Finally, in Sect. 7, we present a brief summary and conclusion and outline future research in this area.

2 Related work

This section provides a survey of the related work from recent literature in the areas of trajectory modeling, view-invariant representation and applications of trajectory-based representation and learning. Studies into human psychology have shown the extra-ordinary ability of human beings to recognize object motion even from minimal information system such as Moving Light Displays (MLDs). Such displays are obtained by making a video of moving subjects wearing reflective pads/light bulbs on their body joints in almost dark conditions. In spite of the paucity of information, human observers easily perceive not only motion but also the kind of motion; e.g., walking, running, dancing, cycling, etc. [23]. Based on this understanding, object motion has been an important feature for the representation and discrimination of one object from another in video applications. Earlier approaches in motion-based methods focused on object tracking from raw and compressed domain videos [16, 21, 37, 38]. Indexing and searching based on object motion as the dominant cue has attracted a lot of research activity in the past few years [2, 8, 11, 13, 14, 39].

View-invariant representation has also been addressed in [35] for modeling and recognizing actions performed by individuals in video sequences. The representation is based on dynamic instants (segmentation points) of the trajectories. For each dynamic instant in the trajectory, frame number, location of the hand and "sign" of the instant ($-$ve for counter clockwise turn and $+$ve for clockwise turn) is stored. The matching is performed on trajectories with the same number of dynamic instants and same sign permutations. This approach though compact in representation is suboptimal and too much dependent on the segmentation process. Also, it has no room for partial trajectory processing. The system is tested on 47 instants of 16 different actions, so it is hard to judge the scalability issues of this system for large datasets. In our previous work on view-invariant representation, we addressed the problem of indexing and retrieval. We used the Fourier descriptor (FD) computed from centroid distance function [3], as well as curvature scale space [4]. One important contribution of our work is that we have formulated the view-invariant trajectory representation as open-ended shape representation problem. Specifically, we argue that the only difference between an image shape and the trajectory of an object is that the former is a closed contour while the later is not. This allows us to take advantage of a wealth of recent work involving affine-invariant image shape description. A number of shape representation schemes for image analysis have been proposed under the affine transform setting [31, 42, 27].

Semantics-based processing of trajectory data to extract high-level information, such as activity recognition, has gained interest quite recently [7, 29, 36]. Dagtas et al. [15] propose several motion description schemes that serve as intermediate spatiotemporal models for event-based retrieval of video. They present trajectory-based and trail-based models for motion-based indexing of videos. They provide the implementation details of PICTURESQUE, a video database retrieval system using a query-by-example framework. In [20], the issue of recognizing a set of plays from American football videos is considered. Using a set of classes each representing a particular game plan and computation of perceptual features from trajectories, the propagation of uncertainty paradigm is implemented using automatically generated Bayesian network. The problem with above approaches is that they are highly domain-dependant, with domain knowledge and sensor dependence on video data being intimately woven into the systems. On similar lines, Nevatia et al. [19] have addressed the issue of activity recognition in single or multiple actor situations which exhibit some specific patterns of whole body motion. Nuria Oliver et al [30] present a system based on Coupled Hidden Markov Models (CHMM) for classifying the types of interactions between humans in a video surveillance setting. Yuri Ivanov et al. [22] address a similar problem using a stochastic context-free grammar parsing mechanism. Stauffer et al. [25] have developed an indexing and retrieval system for video footage in surveillance systems. Their prototype system, Spot, can answer natural language questions about moving objects that appear in a surveillance video footage, like "Did any cars leave the garage towards North." In our previous work on high level trajectory modeling for activity recognition, we have proposed a semisupervised learning approach based on Gaussian Mixture Model (GMM) [5]. In [6] we showed that although the GMM-based approach robustly captures the underlying complex statistical distribution of trajectory data, it fails to model the dynamic time warping aspect of it. For this purpose, we proposed a semisupervised HMM-based trajectory classifier for activity recognition. In the forthcoming sections, we present our view-invariant representation of the trajectories using CSS and CDF-based representation along with HMM-based classifier for activity recognition.

3 View-invariant representation

Our ultimate goal in this paper is matching trajectories and classifying them across different views of similar activities. Trajectories of two objects captured from different viewpoints appear entirely different in terms of their raw data representation. Thus, we wish to use a representation which is invariant to perspective transformation. It is well known in the context of shape representation that a perspective transformation can be approximated by an affine transformation as long as the imaged object is planar and the camera optical center is far enough from the image plane [17]. An object trajectory is essentially a 2-D signal having x- and y-projections of object centroid at each successive frame of video clip. Hence, a trajectory can be modeled as a parametric curve representing the object's x- and y-locations over successive frames as

$$r[k] = \{x[k],\ y[k]\}, \quad k = 0,\ 1, \ldots,\ N-1. \tag{1}$$

In case of shape representation, the above parametric curve is a closed contour specifying object boundary. For the sake of trajectory representation, the parametric curve may not necessarily be closed and the assumption of periodicity implicitly used by some shape representation techniques does not hold. The general affine transformation applied to a trajectory can be formulated mathematically as

$$\begin{gathered} r_{\mathrm{a}}(t) = A_{\mathrm{shear}} A_{\mathrm{rotate}} r(t) + A_{\mathrm{translate}} \\ \begin{bmatrix} x_{\mathrm{a}}(t) \\ y_{\mathrm{a}}(t) \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \cdot \begin{bmatrix} x(t) \\ y(t) \end{bmatrix} + \begin{bmatrix} e \\ f \end{bmatrix} \end{gathered} \tag{2}$$

where $r_{\mathrm{a}}(t)$ is the affine transformed trajectory; A_{shear}, A_{rotate}, and $A_{\mathrm{translate}}$ represent the shear, rotation and translation matrices respectively; and $x_{\mathrm{a}}(t)$ and $y_{\mathrm{a}}(t)$ represent the x- and y-projections of the transformed trajectory. As an example, we provide a small snapshot of the ASL dataset we are using in this presentation. Figure 1 depicts one representative trajectory per class for three different classes in this dataset, along with their rotated versions. As evident from this figure, the effect of rotation is that the raw representation of the rotated version of the trajectory is entirely different for the same trajectory. In this section, we study two feature spaces (CSS and CFD) for shape description in affine invariant settings. These features spaces allow affine representation of trajectories in the form of its constituent subtrajectories thus facilitating the use of powerful learning techniques based on HMMs. Section 3.1 outlines the curvature

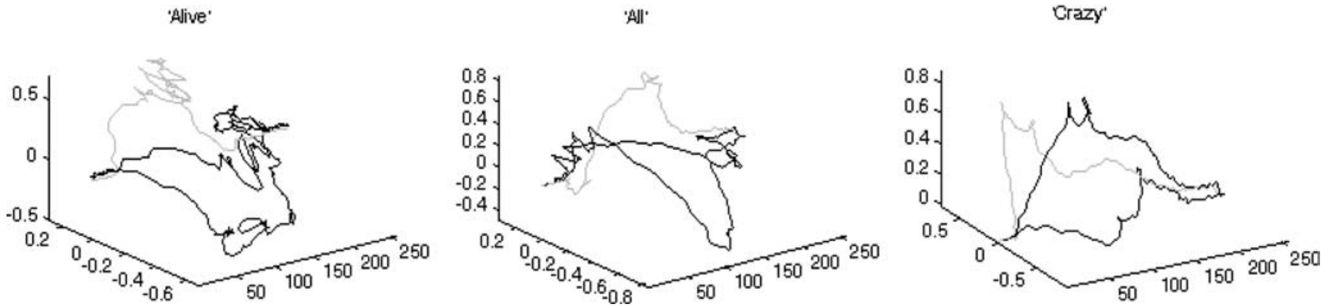

Fig. 1 Three sample trajectories from ASL dataset. For each trajectory, its two rotated version are also shown. Original trajectories are drawn in black

scale space for trajectory representation, while Sect. 3.2 describes the centroid distance function (CDF) representation. Both of these methods yield affine invariant representations.

3.1 Curvature scale space

Scale-space is a multi-resolution technique used to represent data of arbitrary dimension without any knowledge of noise level and preferred scale (smoothness). The notion of scale in measured data is handled by representing measured signal at multiple levels of detail, from the finest (original signal) to the coarsest (most-smoothed version). The CSS representation takes curvature data of a parametric curve and represents it by its different evolved versions at increasing levels of smoothness. The curvature $\kappa[k]$ for the trajectory represented as in Eq. (1) can be expressed as

$$\kappa[k] = \frac{x'[k]y''[k] - y'[k]x''[k]}{\{x'[k]^2 + y'[k]^2\}^{3/2}} \tag{3}$$

The curvature of a trajectory has several desirable computational and perceptual characteristics. One such property is that it is invariant under planar rotation and translation of the curve. Curvature is computed from dot and cross products of parametric derivatives and these are purely local quantities, hence independent of rotations and translations. The dot and cross products are based only on the lengths of, and angles between, vectors. Hence, these are also independent of rigid transformations of the coordinate system.

Given a trajectory as in Eq. (1), the evolved version of the trajectory in terms of scale-space is defined by

$$r_\sigma[k] = \{X[k,\sigma], Y[k,\sigma]\} \tag{4}$$

where

$$\begin{aligned} X[k;\sigma] &= x[k] \otimes g[k;\sigma] \\ Y[k;\sigma] &= y[k] \otimes g[k;\sigma] \end{aligned} \tag{5}$$

with $g[k;\sigma]$ being the symmetric Gaussian kernel used for smoothing. At each level of scale, governed by the increasing standard deviation of Gaussian kernel, curvature of the evolved trajectory is computed. Then the function implicitly defined by

$$\kappa[k;\sigma] = 0 \tag{6}$$

is the curvature scale space image of the trajectory. It is defined as a binary image with a value of 1 assigned to the points of curvature zero crossing at each scale level. Figure 2 depicts an example trajectory from “Alive” class in the ASL dataset along with its three rotated versions. The corresponding CSS images are also shown in the third column of the figure. Note that each of the arch-shaped contours in the CSS image corresponds to a convexity or concavity on the original trajectory with the size of the arch being proportional to the size of the corresponding feature. The CSS image is a very robust representation under the presence of noise in trajectory data due to small camera motions and minor jitters in tracking. Noise amplifies only the small peaks, with no effect on the location or scale of the feature contour maxima [28]. As evident from the figure, the major peaks of the CSS image remain quite preserved after significant affine transformation.

3.2 Centroid distance function

The centroid distance function is another invariant representation of the raw shape data used in the affine invariant image

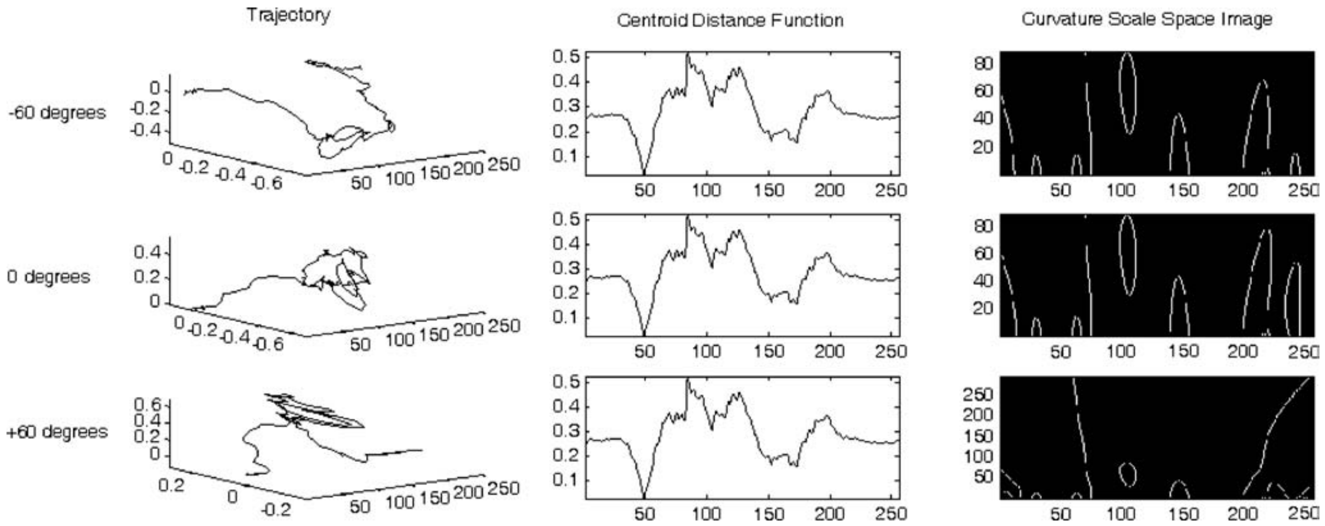

Fig. 2 Trajectory of the hand motion for signing the word “Alive” in Australian Sign Language with its two rotated versions, and their corresponding representations using centroid distance functions (CDFs) and curvature scale-space (CSS) images

retrieval applications. The centroid distance function is expressed by the distance of each point in trajectory from the centroid of the trajectory:

$$c[k] = \left\{\sqrt{[x[k] - x_c]^2 + [y[k] - y_c]^2}\right\}, \quad k = 0, 1, \ldots, N-1 \tag{7}$$
$$\text{where } x_c = \frac{1}{N}\sum_{t=0}^{N-1} x[k], \quad y_c = \frac{1}{N}\sum_{t=0}^{N-1} y[k]$$

Let us denote by *uniform affine transformation* the set of affine transforms including translation, rotation and uniform scaling. This excludes the shear transformation in general affine transformation. Let us denote the centroid distance function of a trajectory before affine transformation as $C\,[K]$ and after uniform affine transformation as $C'\;[K]$. Then it can be easily proved that under uniform affine transformation, the following relation holds between the centroid distance functions computed from original and affined version of the trajectory:

$$C'[k] = \alpha C[k]$$
$$\alpha = 1 \text{ for } \begin{bmatrix} u[k] \\ v[k] \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x[k] \\ y[k] \end{bmatrix} + \begin{bmatrix} \beta \\ \gamma \end{bmatrix}$$
$$\alpha = \alpha_s \text{ for } \begin{bmatrix} u[k] \\ v[k] \end{bmatrix} = \alpha_s \begin{bmatrix} x[k] \\ y[k] \end{bmatrix} \tag{8}$$

Here, the first case combines rotation and translation, while the second case stands for scaling. We note that the effect of scaling can be easily taken out by resampling the trajectory data to a common sample size. This observation is motivated by the fact that the "shape" of the trajectory conveys a lot more information than the "speed" that the activity is performed at. Similarly, the translation can be taken care of by normalizing the x- and y-coordinates to a common range of values. This is again motivated by the fact that the "shape" of the trajectory conveys more important information than the absolute coordinates region of the world coordinate system in which the activity is performed. This means that the most important affine deformation to take care of in the view-invariant settings is the rotation. Figure 2 displays one of the trajectories from ASL dataset along with two of its rotated versions, and its representation in terms of the two feature spaces explored in this presentation. As seen from this figure, the CDF-based representation is absolutely invariant to rotational deformations. The CSS-based representation, on the other hand, results in the curvature zero crossings being consistent but the CSS maxima tend to shift around. In the next section, we outline the procedure for HMM training and classification based on these to invariant feature spaces.

4 HMM-based view-invariant recognition

Model-based recognition and classification has been extensively used in applications such as sign language recognition, action recognition [7], sports video analysis [41], speech/speaker recognition [12], etc. Hidden Markov models have been used successfully in speech recognition [33], shape representation [9], gesture recognition [40], sports video structure analysis [41], etc. In this section, we build on the two view-invariant feature spaces presented in the previous section for modeling different classes of object motion patterns. Specifically we use the HMM-based modeling, wherein the multimodal PDF of each hidden state is represented using Gaussian mixtures. The underlying feature spaces that we use are based on the invariant feature spaces presented in Sects. 3.1 and 3.2. For CSS-based representation, we model the trajectory as a sequence of the salient CSS maxima locations. This representation keeps the temporal nature of the trajectory data intact for HMM-based semisupervised learning performed in the next stage. In case of CDF-based representation, the trajectory data is segmented into subtrajectories and the PCA coefficients of these subtrajectories are then used for compact representation in a view-invariant setting. The individual activity classes are then modeled by HMMs trained from the temporal sequence of subtrajectories in the PCA subspace. Hidden Markov Models are finite state stochastic machines that allow dynamic time-warping for modeling time series data which satisfies the Markovian property. Simply put, the 1st-order Markovian property assumes the independence of the current state from all past states given the previous state. Object trajectory data is a stochastic process with temporal continuity, just like speech signals, which has been successfully modeled using HMMs for the past several decades [33].

Raw trajectory data, however, has one inherent problem that could potentially fail HMM-based modeling: namely, it does not necessarily satisfy the 1st-order Markovian property. For instance, when an object starts moving in a particular direction, it goes along with the flow for a while before changing direction. Therefore, successive trajectory points used to model the states are bound to violate the 1st-order Markovian property of the resulting Markov chain. This point, however, is precisely where the segments-based approaches of trajectory modeling come into picture. In the CSS, as well as CDF-based representation schemes, the trajectory is represented by small segments of constant motion activity whose boundaries are delineated by significant changes in some parameters of motion activity. Our segmented trajectory-based approaches in this context model the trajectory data along exactly the same lines as speech signals. Specifically, we are interested in modeling a class of object motions (words) based on the temporal ordering of subtrajectories (phonemes). In this model, a subtrajectory can be used to model the state of the HMM. Since subtrajectories represent segments of atomic motions between points of change in motion pattern, the resulting Markov chain can be modeled as a 1st-order Markov process. We also observe that mixture of Gaussians is a robust method for estimating the PDF in the absence of dynamic time-warping phenomenon. We therefore propose to use continuous density HMMs, where each state of the HMM is modeled by a mixture of Gaussians. A major problem in HMM design is the topology (left-to-right or ergodic), and the number of states in the HMM. We propose a data-driven design of HMM with no restriction of HMM topology and number of states. In the following subsections, we outline the process of initializing and estimating the parameters of HMMs as well as the classification process. In Sect. 4.1, we design the HMM-based view-invariant recognition system using the CSS-based representation. Section 4.2 outlines the view-invariant HMM-based recognition using CDF-based representation.

4.1 CSS-based representation for activity recognition

In the case of CSS representation, as outlined in Sect. 3.1, the trajectory is represented by the locations of CSS maxima in terms of their temporal ordering and the scale of concavity. In this context, the trajectory data is represented by a time-ordered sequence of two-dimensional feature vectors containing CSS maxima. Here each maxima location corresponds to a concavity in the "shape" of trajectory. Our trajectory modeling scheme represents the class of each activity based on the temporal sequence of these concavities in the CSS image of trajectory. Broadly speaking, our semisupervised learning scheme takes the training set data for each class, computes the CSS images of all the training set instants and models each class using the trained HMM. At the time of query, the CSS maxima of query image are computed and the query feature vector is posed to all the trained HMMs from all classes. The classification result is declared to be the class represented by HMM which results in highest log-likelihood. The following subsections detail the process of training the HMM and classification based on this training.

4.1.1 HMM training and parameter estimation

In terms of training the HMMs, the first parameter specified for an HMM is the number of states. For each class, represented by a separate HMM, we set the number of states to be equal to the number of CSS maxima. Once the number of states is fixed, the complete set of model parameters describing the HMM are given by the triplet

$$\lambda = \{\pi_j, a_{ij}, b_j\} \tag{9}$$

where π_j is the probability of the *j*th CSS maximum being the first maximum among all the maxima, a_{ij} denotes the probability of the *j*th CSS maximum occurring immediately after the *i*th one, and b_j denotes the PDF of *j*th state. We use Gaussian Mixture-based representation for the state PDF. Once the separate HMMs are trained for all classes, recognition of new trajectories can be performed based on the likelihood computed for such trajectories posed to individual HMMs. The following subsections describe this process in detail.

Once the CSS images from the training set of trajectories for a class are computed and the number of states decided, the HMM's parameter triplet in Eq. (9) can be estimated. For a given trajectory, let there be T CSS maxima. Then the state variable q_t which corresponds to the *t*th maximum, takes one of N values $q_t \in \{S_1, \ldots, S_N\}$. Since the Markovian assumption is valid, the probability distribution of q_{t+1} depends only on q_t. This is described by the state transition probability matrix A whose elements a_{ij} represent the probability that q_{t+1} corresponds to state S_j given that q_t corresponds to S_i. The initial state probabilities are denoted by π_i, the probability that q_1 equals S_i. The observational data O_t from each state of the HMM is generated according to a PDF dependent on the state at the instant of *t*th CSS maximum, denoted by $b_j(O_t)$. This state-conditional observation PDF is modeled as a Gaussian mixture given by

$$b_j(O_t) = \sum_{m=1}^{M} c_{jm}\mathbb{N}(\mu_{jm}, \Sigma_{jm}) = \sum_{m=1}^{M} c_{jm}\frac{1}{(2\pi)^{P/2}\left|\Sigma_{jm}\right|^{1/2}} \times \exp\left\{-\frac{1}{2}(O-\mu_{jm})^{\mathrm{T}}\Sigma_{jm}^{-1}(O-\mu_{jm})\right\} \tag{10}$$

where c_{jm}, μ_{jm} and Σ_{jm} denote the scalar mixing parameter, P-dimensional mean vector and $P\times P$ covariance matrix of *m*th Gaussian component in *j*th state. Here, each Gaussian component is a multivariate normal distribution of the same dimensionality as the CSS maxima feature vector, namely two-dimensional. The parameters of the HMM are initialized to random values and the Baum–Welch algorithm is used to re-estimate the parameters using the forward-backward procedure [33]. The above discussion relates to training a sequence of subtrajectories resulting from one trajectory. Given a set of trajectories corresponding to each class, we extend the training to multiple training set trajectories. At each iteration of the Baum-Welch re-estimation, the contribution from all of the individual training set trajectories are summed up in the forward-backward estimation parameters. Once the change in parameter values is less than a prefixed threshold for 10 successive iterations, the algorithm is declared to have converged. One problem with this form of parameter estimation is that it can get stuck in a local minimum no matter how many iterations are performed. To improve the performance in this situation, we use the process of *annealing*. After one set of iterations for parameter estimation, we perform the annealing by expanding the covariance matrices of PDF estimates and by pushing the state transition matrix and prior state probabilities closer to "uniform." This has the effect of increasing the "temperature"; thereby, nudging the solution away from local peaks in search of a global maximum.

4.1.2 Trajectory classification

Once the HMMs for all classes have been trained, the classification of new trajectories can be performed by computing the likelihoods. For this purpose, the 2-D feature vector of CSS maxima extracted from input trajectories are posed as an observation sequence to each HMM. Given HMMs for the L classes, $\lambda_1, \lambda_2, \ldots, \lambda_L$, and the set of CSS maxima feature vectors of input trajectory $Y_1, Y_2, \ldots, Y_m$, it is declared to belong to the class represented by the HMM with the highest likelihood. Given the set of CSS maxima feature vectors of the input trajectory, the probabilities $p(Y_1, Y_2, \ldots, Y_m | \lambda_i)$ are computed for all classes using the forward-backward procedure. The decision rule then becomes:

$$\text{class} = \arg \max_{i \in [1,\ldots,L]} p(Y_1, Y_2, \ldots, Y_m | \lambda_i) \tag{11}$$

The results of this training and classification process on CSS image maxima are presented in Sect. 5.

4.2 CDF-based representation for activity recognition

This section provides a very brief overview of our trajectory-modeling scheme based on CDF-based trajectory representation. We recognize that most often full trajectory information is unavailable in video tracking applications due to occlusions. This limitation requires trajectory representation methods that can perform well even in the case of partial trajectory information. We address this problem by segmenting the trajectories at the points of perceptual discontinuities in their CDF representation. In order to reduce the dimensionality of the feature space, each subtrajectory is represented by a set of its PCA coefficients. The HMMs are then trained for trajectory classification and activity recognition using their PCA coefficients.

Our subtrajectory-based representation has several advantages: Firstly, it is motivated by motion perception in humans which is highly dependent on piecewise segments based on atomic units of actions [23]. Secondly, it facilitates the modeling and recognition of trajectories which only have partial trajectory information available. And, finally, this process decomposes the large trajectory data to a few temporally-sequenced observation vectors which obey the Markovian property. For the purpose of segmentation, the discontinuities in the trajectory are detected with the help of velocity (1st-derivative) and acceleration (2nd-derivative). From the x and y-projections of the trajectory data, we compute the curvature which measures the sharpness of a bend in a 2-D curve and captures derivatives up to 2nd-order as given by Eq. (3). We perform a hypothesis testing-based process to locate the points of maximum change of the curvature data. These inflection points are detected using a likelihood ratio test-based approach. More details can be found in [3].

4.2.1 PCA for subtrajectory representation

We represent the subtrajectories obtained from the 1-D CDF representation using PCA because of its optimal energy compaction properties resulting from custom bases derived from the data [24]. All the vectors of trajectories from all the classes are stacked to form one data matrix. The principal components of this data matrix are then estimated using Eigenspace decomposition of the estimated covariance matrix [24]. To achieve dimensionality reduction, only the first M principal components (PCs) are retained to form the transformation matrix Φ_M. The pool of subtrajectories is finally represented by their PCA coefficients using the transformation:

$$Y = \Phi_M^T [X - \bar{X}] \tag{12}$$

where X denotes the data matrix of subtrajectories, $\bar{X}$ is the vector containing the mean of the data set, and Y is the matrix containing the PCA coefficients of all subtrajectories. The set of PCA coefficients of all subtrajectories for each class are subsequently used to train an HMM for each class as explained in the next subsection.

4.2.2 HMM training and classification

The HMM training and classification process is quite similar to the outline of Sect. 4.1.1 except with a few differences described here. The number of states for each class is set to be equal to the number of subtrajectories. Once the number of states is fixed, the complete set of model parameters describing the HMM are given by the triplet:

$$\lambda = \{\pi_j, a_{ij}, b_j\} \tag{13}$$

where π_j is the probability of the jth subtrajectory being the first subtrajectory among all the trajectories, a_{ij} denotes the probability of the jth subtrajectory occurring immediately after the ith subtrajectory, and b_j denotes the PDF of jth state. We use Gaussian Mixture-based representation for the state PDF.

Once the set of training trajectories for a class are segmented and the number of states decided, the HMM's parameter triplet in Eq. (13) can be estimated. For a given trajectory, let there be T subtrajectories. Then the state variable q_t which corresponds to the tth subtrajectory, takes one of N values $q_t \in \{S_1, \ldots, S_N\}$. Since the Markovian assumption is valid, the probability distribution of q_{t+1} depends only on q_t. This is described by the state transition probability matrix A whose elements a_{ij} represent the probability that q_{t+1} corresponds to state S_j given that q_t corresponds to S_i. The initial state probabilities are denoted by π_i, the probability that q_1 equals S_i. The observational data O_t from each state of the HMM is generated according to a PDF dependent on the state at the instant of tth subtrajectory, denoted by $b_j(O_t)$. This state-conditional observation PDF is modeled as a Gaussian mixture given by

$$b_j(O_t) = \sum_{m=1}^{M} c_{jm}\mathbb{N}(\mu_{jm}, \Sigma_{jm}) = \sum_{m=1}^{M} c_{jm}\frac{1}{(2\pi)^{P/2}\left|\Sigma_{jm}\right|^{1/2}} \times \exp\left\{-\frac{1}{2}(O-\mu_{jm})^{\mathrm{T}}\sum_{jm}^{-1}(O-\mu_{jm})\right\} \qquad (14)$$

where c_{jm}, μ_{jm} and Σ_{jm} denote the scalar mixing parameter, P-dimensional mean vector and $P \times P$ covariance matrix of mth Gaussian component in jth state. Here, each Gaussian component is a multivariate normal distribution of the same dimensionality as the PCA coefficients representing the subtrajectories. More details on computing the parameters of Gaussian mixtures can be found in [5]. Once all the HMMs are trained with one HMM per class, recognition of new trajectories can be performed based on the likelihood computed for such trajectories posed to individual HMMs. This process is similar to the description in 0.

5 Experimental analysis and performance evaluation

This section analyzes the performance of our CDF and CSS-based representations for view-invariant activity recognition. In the following subsections, we present the details of the dataset and then the procedure and results of experiments.

5.1 Data set

In this section, we outline the details of the data set used in our experiments. We use the Australian Sign Language (ASL) data set, obtained from UCI's KDD archives. These trajectories are obtained by registration of the hand coordinates at each successive instant of time by using a Power Glove interfaced to the system. The system registers 3-D positions of the hand, palm orientation, and several other features at each sampling instant as five professional signers sign around 95 words in multiple sessions. For each word, there are 69 recordings of signing activity across all signers. Out of these set of trajectories, we extract the x and y-locations corresponding to five classes for training and recognition in the view-invariant setting. For the semisupervised training part, we use 50% of the original trajectories from each class. This corresponds to a training set of $34 \times 5 = 170$ trajectories. To perform testing in the view-invariant setting, we generated a test set by rotating all the trajectories in the original dataset. The set of rotations to generate affine variations consisted of 5 angles $\{-60, -30, 0, 30, 60\}$. Thus the test dataset consists of $69 \times 5 \times 5 = 1725$ trajectories. The data set shows a lot of variation in trajectory data for the same words and even for the same signer. The spatial and temporal variation arises from various sources including: noise in the sensors, variation in skill levels among different signers, and fatigue on signers performing the task.

5.2 Simulation results

This section summarizes the results of our computer simulation experiments. We report the performance of the

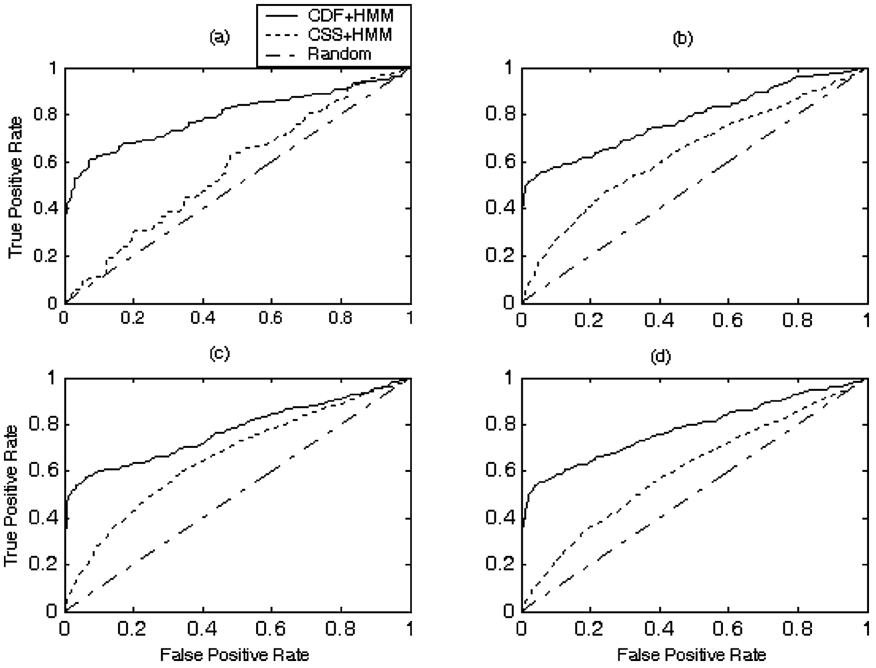

Fig. 3 ROC curves for the rotated trajectories posed for classification. Number of classes: **a** 2, **b** 3, **c** 4, **d** 5

HMM-based classification system using the two feature spaces in terms of ROC curves and accuracy for a test data set of 1725 trajectories with five classes. We compare the performance of the two HMM-based systems among themselves and with the random classifier. The random classifier is obtained by labeling all the query trajectories with randomly generated class labels and then comparing with the ground truth. We report the results on a range of dataset sizes from the ASL dataset. The results are reported in terms of the ROC curves in Fig. 3. The ROC curves are two-dimensional depiction of classifier performance. To compare classifiers, we may want to base our quantitative analysis on a single scalar value. The probability of accuracy is such a scalar quantity that represents classifier performance tradeoff. This result is provided for different dataset sizes from the ASL dataset in Fig. 4.

Based on these results, we note that the CDF-based approach yields superior recognition results. This performance difference can be attributed to a number of facts already highlighted in the text. Firstly, the CDF-based representation is invariant to affine transformation without regards to level of smoothing performed on the signal. On the other hand, the CSS-based representation tends to generate spurious maxima at higher levels of smoothing. This can be seen in Fig. 2 where the CDF-based representation is shown to be truly invariant to rotational transformation, while the CSS-based representation suffers from distortions at higher levels of smoothing. Secondly, our CDF-based approach represents the trajectories using PCA coefficients of constituent subtrajectories, as opposed to only the points of CSS maxima in the case of CSS-based representation. This robustness in representation based on the subtrajectory model results in the higher performance of the CDF-based representation. From Fig. 4 we also note that the relative accuracy of the CDF-based HMM classifier compared with CSS-based representation, increases with an increase in the number of classes; thus making it more scalable for large number of classes. Moreover, we note that much higher probability of accuracy values would have been attained provided the size of the training data sets would have been increased proportionally. Furthermore, the HMM-based trajectory modeling system proposed in this paper where individual states are modeled by mixtures of Gaussians has been shown to perform consistently better than the other trajectory modeling techniques used in all of our experiments.

6 Summary and conclusions

In this paper, we have presented a detailed discussion on the topic of motion trajectory-based statistical modeling and classification of data in a view-invariant setting. Our aim in this presentation has been to motivate the need for, and challenges involved in, the classification and recognition of temporal data resulting from object tracking captured through multiple views. We cast the view-invariant activity recognition problem as affine-invariant image shape retrieval. This approach detects the CSS contour maxima and represents the trajectories based on the locations of these peaks. The limitation of this approach is that it does not model the data between segmentation points. This shortcoming is alleviated in the CDF+PCA based representation scheme. Here, the centroid distance function from the trajectory data is computed. The CDF-based representation is used to segment the trajectories and the PCA coefficients of the segmented data are then stored to train the HMMs on. We have based our experiments on ROC curves as well as accuracy values. The classification systems were tested on a standard ASL data set. The results are generated for various sizes from 690 to 1725 trajectories in this dataset.

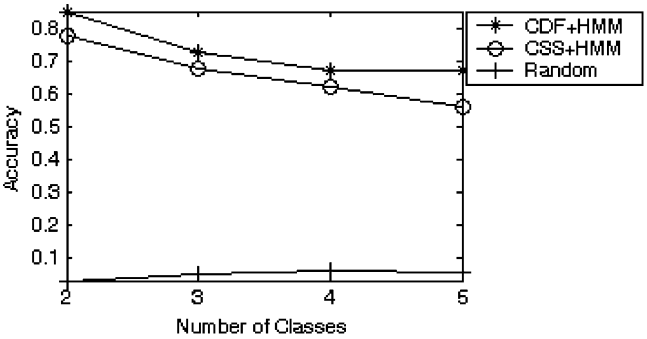

Fig. 4 Accuracy values for the classification systems on increasing number of classes

Future research must focus on motion trajectory-based modeling and classification of video sequences that are robust to camera orientation and movement. Generalization of the proposed PCA representation to nonlinear transformations (e.g., Kernel PCA or Kernel Discriminant Analysis), are needed to deal with nonlinear classification. An important extension of our approach would be required to perform multiple motion trajectory-based classification for "semantic" retrieval from video sequences. It is also possible that the basis of our approach could be used for video sequence mining by detection and identification of motion trajectories in the video query.

References

1. Bashir, F., Khanvilkar, S., Schonfeld, D., Khokhar, A.: Multimedia systems: content-based indexing and retrieval. In: Chen, W.K. (ed.) The Electrical Engineering Handbook, Sect. 4, Chapter 6. Academic Press (2004)
2. Bashir, F., Khokhar, A., Schonfeld, D.: Segmented trajectory based indexing and retrieval of video data. In: International Conference on Image Processing. Barcelona, Spain (2003)
3. Bashir, F., Khokhar, A., Schonfeld, D.: A hybrid system for affine-invariant trajectory retrieval. ACM SIGMM Multimedia Information Retrieval Workshop, New York, NY (2004)
4. Bashir, F., Khokhar, A.: Curvature scale space based affine-invariant trajectory retrieval. In: IEEE International Multitopic Conference, INMIC 2004. Lahore, Pakistan (2004)
5. Bashir, F., Khokhar, A., Schonfeld, D.: Automatic object trajectory-based motion recognition using gaussian mixture models. In: IEEE International Conference on Multimedia & Expo (ICME 2005). Amsterdam, the Netherlands (2005)

6. Bashir, F., Qu, W., Khokhar, A., Schonfeld, D.: HMM-based motion recognition system using segmented PCA. In: IEEE International Conference on Image Processing (ICIP 2005). Genoa, Italy (2005)
7. Brand, M., Oliver, N., Pentland, A.: Coupled hidden markov models for complex action recognition. In: Proceedings Conference on Computer Vision and Pattern Recognition, p. 994 (1997)
8. Buzan, D., Sclaroff, S., Kollios, G.: Extraction and clustering of motion trajectories in video. In: International Conference on Pattern Recognition (2004)
9. Caelli, T., McCabe, A., Briscoe, G.: Shape tracking and production using hidden markov models. Int. J. Pattern Recognit. Artificial Intell. **15**(1), 197–221 (2001)
10. Chang, S.F., Chen, W., Meng, H.J., Sundaram, H., Zhong, D.: A fully automated content-based video search engine supporting spatiotemporal queries. IEEE Trans. Circ. Sys. Video Techn. **8**(5) (1998)
11. Chen, L., Ozsu, M.T., Oria, V.: Symbolic representation and retrieval of moving object trajectories. ACM SIGMM Multimedia Information Retrieval Workshop. New York (2004)
12. Chen, T., Huang, C., Chang, C., Wang, J.: On the Use of Gaussian Mixture Model for Speaker Variability Analysis. ICSLP. Denver, Colorado (2002)
13. Chen, W., Chang, S.F.: Motion Trajectory Matching of Video Objects. SPIE. San Jose, CA (2000)
14. Cheung, S., Zakhor, A.: Fast similarity search on video sequences. In: Proceedings IEEE International Conference on Image Processing (2003)
15. Dagtas, S., Al-Khatib, W., Ghafoor, A., Kashyap, R.: Models for motion-based video indexing and retrieval. IEEE Trans. Image Process. **9**(1), 88–101 (2000)
16. Dimitrova, N., Golshani, F.: Motion recovery for video content classification. ACM Trans. Inf. Syst. **13**(4), 408–439
17. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, Cambridge (2003)
18. Hettich, S., Bay, S.D.: The UCI KDD Archive [http:///kdd.ics.uci.edu]. University of California, Department of Information and Computer Science, Irvine, California (1999)
19. Hongeng, S., Nevatia, R., Bremond, F.: Video-based event recognition: Activity representation and probabilistic recognition methods. Comput. Vis. Image Understanding **96**, 129–162 (2004)
20. Intille, S.S., Bobick, A.F.: Recognizing planned, multiperson action. Comput. Vis. Image Understanding **81**, 414–445 (2001)
21. Isard, M., Blake, A.: A Mixed-State CONDENSATION tracker with automatic model-switching. In: Proceedings of the International Conference on Computer Vision, pp. 107–112 (1998)
22. Ivanov, Y.A., Bobick, A.F.: Recognition of visual activities and interactions by stochastic parsing. IEEE Trans. Pattern Anal. Machine Intell. **22**(8), 852–872 (2000)
23. Johansson, G.: Visual perception of biological motion and a model for its analysis. Percept. Psychophys. **14**(2), 201–211 (1973)
24. Jolliffe, I.T.: Principal Component Analysis. Springer-Verlag, New York (1986)
25. Katz, B., Lin, J., Stauffer, C., Grimson, E.: Answering questions about moving objects in surveillance videos. In: Proceedings of AAAI Spring Symposium on New Directions in Question Answering (2003)
26. Moghaddam, B., Wahid, W., Pentland, A.: Beyond EigenFaces: probabilistic matching for face recognition. In: International Conference on Automatic Face and Gesture Recognition. Nara, Japan (1998)
27. Mokhtarian, F., Abbasi, S.: Retrieval of similar shapes under affine transformation. In: Proceedings of the International Conference on Visual Information Systems. Amsterdam, The Netherlands, pp. 566–574 (1999)
28. Mokhtarian, F., Bober, M.: Curvature Scale Space Representation: Theory, Applications, and MPEG-7 Standardization. Kluwer Academic Publishers, Netherlands (2003)
29. Naphade, M., Kozintsev, I., Huang, T.: Factor graph framework for semantic video indexing. IEEE Trans. Circuits Syst. Video Technol. **12**(1) (2002)
30. Oliver, N.M., Rosario, B., Pentland, A.: A bayesian computer vision system for modeling human interactions. IEEE Trans. Pattern Anal Machine Intell. **22**(8), 831–843
31. Pentland, A., Sclaroff, S.: Modal matching for correspondence and recognition. IEEE Trans. Pattern Anal Machine Intell. **17**(6), 545–561 (1995)
32. Porikli, F.M.: Trajectory distance metric using hidden markov model based representation. In: Sixth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, PETS (2004)
33. Rabiner, L.R.: A tutorial on Hidden Markov Models and selected applications in speech recognition. In: Proceedings of the IEEE, vol. 77, pp. 257–286 (1989)
34. Rangarajan, K., Allen, W., Shah, M.: Matching motion trajectories using scale-space. Pattern Recognit. **26**(4), 595–610 (1993)
35. Rao, C., Yilmaz, A., Shah, M.: View-invariant representation and recognition of actions. Int. J. Comput. Vis. **50**(2), 203–226 (2002)
36. Rea, N., Dahyot, R., Kokaram, A.: Semantic event detection in sports through motion understanding. In: Proceedings of Conference on Image and Video Retrieval. Dublin, Ireland (2004)
37. Sahouria, E., Zakhor, A.: A Trajectory based video indexing system for street surveillance. In: IEEE International Conference on Image Processing (1999)
38. Schonfeld, D., Lelescu, D.: VORTEX: Video retrieval and tracking from compressed multimedia databases—multiple object tracking from MPEG-2 bitstream (Invited Paper). J. Vis. Commun. Image Representation **11**, 154–182 (2000)
39. Vlachos, M., Kollios, G., Gunopulos, D.: Discovering similar multidimensional trajectories. In: Proceedings of the International Conference on Data Engineering, p. 673 (2002)
40. Wilson, A.A., Bobick, A.F.: Hidden Markov Models for modelling and recognizing gesture under variation. Hidden Markov Models: Appl. Comput. Vis. pp. 123–160 (2001)
41. Xie, L., Chang, S.F., Divakaran, A., Sun, H.: Structure analysis of soccer video with Hidden Markov Models. In: IEEE International Conference on Acoustic, Speech and Signal Processing. Orlando, FL (2002)
42. Zhang, D.S.: Image retrieval based on shape. Ph.D Thesis, Monash University, Australia (2003)